APPENDIX

Examples for Docket AM9-97-120 “Filtering of information entities”

October 8, 1997

1. **hypertext pages (e.g. www pages) and hyperlinks (e.g. HREFS):** One potential affinity measure is to let:

$$\begin{aligned} a_1(u,v) &= 1 \text{ if the link } u \to v \text{ exists} \\ &= 0 \text{ otherwise.} \end{aligned}$$

   Another possibility is:

$$\begin{aligned} a_2(u,v) &= 1 \text{ if the link } v \to u \text{ exists} \\ &= 0 \text{ otherwise.} \end{aligned}$$

   In matrix terms, the latter affinity matrix is the *transpose* of the former i.e. $A_2 = A_1^T$. In this case the associated similarity matrices might be, for example:

$$\begin{aligned} M_1 &= A_1 A_1^T \\ M_2 &= A_2 A_2^T \\ M_3 &= \alpha A_1 A_1^T + (1-\alpha) A_2 A_2^T \text{ where } \alpha \in [0,1]. \end{aligned}$$

   In component form:

$$\begin{aligned} m_1(u,v) &= \textstyle\sum_w a_1(u,w) a_1(v,w) \\ m_2(u,v) &= \textstyle\sum_w a_2(u,w) a_2(v,w) \\ m_3(u,v) &= \alpha m1(u,v) + (1-\alpha) m2(u,v) \\ &= \alpha \textstyle\sum_w a_1(u,w) a_1(v,w) + (1-\alpha) \textstyle\sum_w a_2(u,w) a_2(v,w) \text{ where } \alpha \in [0,1]. \end{aligned}$$

   Note that:

   (a) $m_1(u,v)$ is a measure of the number of web pages “pointed to” by both u and v.

   (b) $m_2(u,v)$ is a measure of the number of web pages that “point to” both u and v.

   (c) $m_3(u,v)$ is a weighted combination of m_1 and m_2. Many more affinity measures and similarity definitions are possible.

2. **documents and terms:** A potential affinity measure is the following:

$$\begin{aligned} a(u,v) &= 1 \text{ if document } u \text{ contains term } v \\ &= 0 \text{ otherwise.} \end{aligned}$$

   Other possible affinity measures are:

- $a(u,v) \;=\;$ the number of times term v occurs in document u.
- $a(u,v) \;=\;$ the *relative frequency* of term v in document u. This is equal to the number of occurences of term v divided by the total number of term occurences (all terms) in document u.
- $a(u,v) \;=\;$ the so-called TF/IDF measure of term v in document u, which is the frequency of term v in document u divided by the average frequency of terms v in the entire collection.

Consider the following three-document example:

- Documents:
  (a) the King James Bible
  (b) the novel *Jaws*
  (c) *The Joy of Cooking*
- Terms:
  (a) thou
  (b) shark
  (c) flour
  (d) water

In this case the document-term affinity matrix $A = \{a(u,v)\}$ might look like the following:

$$\begin{bmatrix} 6000 & 10 & 100 & 200 \\ 0 & 3215 & 40 & 3060 \\ 0 & 133 & 3321 & 2856 \end{bmatrix}$$

If we define the document-document similarity matrix to be $M = AA^T$ then $M =$

$$\begin{bmatrix} 36050100 & 648150 & 904630 \\ 648150 & 19701425 & 9299795 \\ 904630 & 9299795 & 19203466 \end{bmatrix}$$

We see from this that, by this measure, *Jaws* and *The Joy of Cooking* are more similar to each other than to The King James Bible because of the frequencies of "water" and to a lesser extent the term "shark".

3. **collaborative filtering example - movie rating:** In this case we have two sets of entities: movies and viewers. The affinity $a(u,v)$ will be a number between 20 and 0 indicating the degree to which viewer u liked or disliked (if less than 10) movie v. Assume viewers Sam, Bill, Ellen, Fred and Mary and the following movies:

   (a) Star Wars
   (b) Die Hard
   (c) My Dinner With Andre

(d) The Rocky Horror Show

(e) Blade Runner

(f) The Remains of the Day

(g) Taxi Driver

(h) Dumb and Dumber

We might, for example, have the following affinity matrix A =

$$\begin{bmatrix} 20 & 20 & 0 & 7 & 17 & 2 & 16 & 10 \\ 18 & 17 & 2 & 10 & 16 & 3 & 19 & 11 \\ 14 & 2 & 17 & 9 & 10 & 19 & 10 & 0 \\ 17 & 19 & 0 & 10 & 17 & 0 & 18 & 20 \\ 18 & 10 & 16 & 14 & 14 & 19 & 12 & 0 \end{bmatrix},$$

which would give the following movie-movie similarity matrix, using $M = A^T A$:

$$\begin{bmatrix} 1533 & 1237 & 562 & 868 & 1309 & 702 & 1324 & 738 \\ 1237 & 1154 & 228 & 658 & 1095 & 319 & 1125 & 767 \\ 562 & 228 & 549 & 397 & 426 & 633 & 400 & 22 \\ 868 & 658 & 397 & 526 & 735 & 481 & 740 & 380 \\ 1309 & 1095 & 426 & 735 & 1130 & 538 & 1150 & 686 \\ 702 & 319 & 633 & 481 & 538 & 735 & 507 & 53 \\ 1324 & 1125 & 400 & 740 & 1150 & 507 & 1185 & 729 \\ 738 & 767 & 22 & 380 & 686 & 53 & 729 & 621 \end{bmatrix}$$

Or, using $M = AA^T$ gives a viewer-viewer similarity matrix M =

$$\begin{bmatrix} 1498 & 1462 & 751 & 1567 & 1126 \\ 1462 & 1464 & 817 & 1563 & 1175 \\ 751 & 817 & 1131 & 716 & 1291 \\ 1567 & 1563 & 716 & 1763 & 1090 \\ 1126 & 1175 & 1291 & 1090 & 1577 \end{bmatrix}$$

Table 3 shows principal and first non-principal affinity components of the movie-movie similarity matrix. The first gives a kind of popularity rating whereas the second shows clustering of movies. In this case the affinity components are the *eigenvectors* of the similarity matrix. The concept of the eigenvectors of a matrix is well known in the theory of linear algebra. They can be computed using any of a number of iterative algorithms like those covered by this invention. The eigenvector concept and a number of iterative algorithms for computing them are described in the book *Matrix Computations* by G. Golub and Charles Van Loan published in 1989 by the Johns Hopkins University Press (ISBN 0-8018-3739-1). Three clusters are evident. At one extreme are *Die Hard* and *Dumb and Dumber*, while at the other are *The Remains of the Day* and *My Dinner With Andre*. The others lie in a somewhat more diffuse central cluster; though *Taxi Driver* and *Blade Runner* are fairly tightly grouped.

The first non-principal affinity component of the viewer-viewer similarity matrix is shown in Table 3, clearly indicating two clusters - men and women in this case.

Star Wars	0.4954	-0.0518
Die Hard	0.4067	0.3172
My Dinner With Andre	0.1733	-0.5587
The Rocky Horror Show	0.2818	-0.1433
Blade Runner	0.4261	0.0486
The Remains of the Day	0.2166	-0.6209
Taxi Driver	0.4332	0.1042
Dumb and Dumber	0.2521	0.4066

Table 1: Principal and first non-principle affinity components (eigenvectors in this case) for movies

Sam	-0.2839
Bill	-0.2147
Ellen	0.6238
Fred	-0.4291
Mary	0.5478

Table 2: First non-principle affinity (eigenvector) components for viewers.